論集第12号（1994年）

多言語対応テキスト検索ツールの自作：ＤＯＳ環境の場合

出　口　厚　実

Two Text Search Utilities for Multilingual MS－DOS System

Atsumi DEGUCHI

Conventional text retrieval systems fall short in one or more way:

*they cannot handle properly textual data containing words which are separated by a hyphen.
*they don't provide a sentence context for the hit word.
*they cannot locate the page of each word's ocurrences.
*they don't allow us to browse freely the search results.

To overcome these shortcomings I have developed the following two text search programs executable on any MS－DOS system. I believe that they are particularly useful for linguists and foreign language teachers or students:

"LFIND.EXE ver. 0.95" is a text search utility which will allow you to search a list of words and provides you with a "key-word-in-sentence" display. It also help you to launch a specific file browser or your favorite editor.

"WFIND.EXE ver. 3.01" will let you scan through a single file, a set of files or a collection of unstructured text and find the actual occurrences of a given word or string fully or partially specified (that is, you can use wildcards for a key word). It gives you the location (the page and line) of all occurrences of your word/string with the sentence in which it is found.

序)

外国語の研究・教育の一部に電子化データを採り入れる気運は最近とみに高まってきている。単に文字情報の処理のみでなく、音声や画像と一体化したいわゆるマルチメディアを支援するハードとソフトの環境が除々に整いつつあることとも関連している。コンピュータの言語研究への応用が始まった初期から注目され、また最も広く行われるようになったのは、やはりテキスト（特に長大なテキストやテキスト資料体）データの電子化とその処理の自動化・高速化である。限られた能力のパーソナルコンピュータを利用する個人中心の作業では、実際、このような文字情報の大量処理こそが現実的で、また成果を得やすい最適な領域であるという事情は現在でも続いていると言ってもよい。

外国語の平文テキストを検索して必要な情報を抽出し、それを加工するためのパッケージソフトはわずかながら存在するものの、使い勝手がよくなく、日常の使用には向かないために、特に筆者の専攻するスペイン語やその他の非英欧文に対応できるようなアプリケーションや小道具類を自前で作ることで何とか対処してきた。前稿「言語研究のための外国語テキストデータ検索：ウィンドウズ環境の場合」（『論集11号』pp.11－30）に紹介した自作ソフト、wwf.exe, wwfi.exe, wsch.exe, dfind.exeなども、その一例である。これらのウィンドウズ用検索ソフトの開発以前に公開していた、いくつかのMS－DOSプログラムを大幅に修正し、多言語対応に改めるとともに、操作性を改善して多種類のテキスト・フォーマットに柔軟に適応できるようにすることにした。これを思い立った理由は、Windowsのユーザが増えたとはいえ、まだまだDOSの利用者も少なくなく、ある程度のテキストデータを蓄積して活用している人でも、わざわざWindows用データに変換するのが面倒である(1)とか、使い慣れたDOSの環境から離れがたいケースもあるほか、データプールから入手できるデータが主にDOS向けに作られていて、それ以外に転用する仕方がわからずに戸惑う初心者も多いという声を耳にしたためである。

本稿で新たに紹介するのは次の2つのテキスト検索ユーティリティで、IBM AT互換機のMS－DOS/V5.0以上、PC－DOS V 6.1以上、NEC98（またはその互換機）のMS－DOS v.3.1以上のいずれでも動作する。また日本語・英語MS－Windows 3.1のDOSモード、OS/2.2.1JのDOS互換ボックスでも使用できるほか、マッキントッシュのSoft PC（そして、恐らくSoft Windowsでも）上のDOS emulationでも問題なく動作する：

“lfind.exe” ver 0.95「多言語対応単語リスト検索プログラム」

2つ以上の単語からなるリスト（1語のみでも可）をone passで検索し、それらを含む「文」を切り出し、該当語をカラー表示する。検索結果を直ちに編集したり、ブラウズするための起動

用アプリケーションを環境変数で指定可能。

“wfind.exe” ver.3.01「多言語・マルチファイル対応文例検索プログラム」

検索対象ファイルを複数併記したり、DOSのワイルドカード表記による一括指定が可能で、グループ化した多数のテキストファイルを一度で検索できる。単語は前方部分一致、後方部分一致条件でも検索でき、それらを含む「文」を切り出し、ファイル名、ページ、行を示す。検索結果を直ちに見たり編集するためのエディタなどを指定できる。

Ⅰ．MS－DOSでのテキスト検索

1．言語データの検索

テキスト検索は何を目的としてなされるかに従い、それに要求される検索機能の重点も変化する。一般に「検索」は汎用の不定形情報として考えられる新聞・雑誌記事のような文章テキストや、各種データベースの中の特定フィールドの中の文字情報（例えば文献カタログ、顧客名簿など）を対象とされることが多い。また、エディタの中での「検索」などのようにプログラムのソースコードという特殊な性格のテキストを想定しているケースもある。

外国語教育・研究の現場で求められている情報のうち、特に需要の多いのはターゲット言語の自然なテキストから得られる特定の語彙・文法形式に付随するコンテキストに関するものである。このような利用目的で言語テキストを一般的なツールで探索しようとするとき、様々な不便が生じる点を前の諸稿で指摘してきた。この稿で問題にする「検索」もこのような外国語の、構造化されない自然な文章の集合体をデータとして、そこから教育や研究に役立つ情報を取り出したい、という用途を考慮した場合を指す。この点では、広義の“データ検索ツール”の汎用性からはずれる点もあるが、しかし、われわれの用途に不可欠な条件をまず解決し、その操作性の改善や機能の充実を目指さなければならないという姿勢を優先した。

特に、自らの経験を中心に既に対応済みの項目も継承して、検索ユーティリティは次のような基本的な事項を押さえるべきと判断した。

(1)

1．検索目標は「語」のみでも十分である。最も使用頻度の高いのは、やはり単語であって、「文字列」ではない。テキスト・ユーティリティとして流通しているほとんどの検索ソフトウェアが文字列をdefaultとして、「語」を二次的な扱いとするのと対照的である。
2．正規表現は（少なくとも欧文の）自然言語テキストの探索に必ずしも威力を発揮しない。
3．出力範囲は「文」が適当である（出口1989,1990bで指摘済み）。
4．所在情報データとして「ページ」は必須である。

5．一連の関連する語群を同時に検索出来ることが望ましい。言語データとして一定範囲に存在する2つ以上の語を探索する場合、より用いる頻度が高いのはOR検索である。例えば、good better bestを同時検索するなど、屈折系列を探索するケースはsyntagmaticな共起関係を調べるAND検索の場合よりも多い。

6．複数テキストファイルの一括指定が可能なほうが便利である。

7．検索の前のpreeditをなるべく軽減できるよう、検索ソフトの側のoption指定の範囲を広めカスタマイズ可能性を高める。すなわち、テキストデータを検索ツールに合わせるという側面は完全には排除できないが、できるだけテキストの形式にツール側を適合できるように配慮する。

8．「語」の判定条件をやや緩くして、諸外国語の区別符号（diacritic）付き文字の代字記号としての機能を考慮に入れる。

9．検索結果を直ちにブラウズでき、また編集作業に迅速に取りかかれることが作業能率に大いに影響する。一時的なスクリーン出力のみで画面から消えた部分が見えなくなってしまうソフトは非常に使いにくく、出力をリダイレクトして書き出したファイルを、わざわざエディタを呼び出して編集するのは2度手間になる。

2．基本性能と汎用性

1つのソフトが各種パソコンの複数のOSでそのまま利用できれば、有り難い。しかし、同一のプログラムを異質なシステムで動作させるのは、本来、不可能であるから、普及度の高いOSの上で機種依存なく幅広く利用できる検索ソフトがより望ましい。この点でMS－DOSはパソコン界ですでに標準的な地位を占め、国内国外を問わず、外国語研究者や教育者の間でも広く普及していると思われる。ただし、前稿の出口（1993、1994b）で取り上げたように、日本語MS－DOSとオリジナルなMS－DOSあるいは欧州各国向けにローカライズされているバージョンとの間に、重大な相違が存在する。後者のDOSは以下で便宜上英語MS－DOSと呼ぶことにする。

再説しないが、テキスト・データの規格そのものにも係るこの重要な差異をどのようにソフト側で吸収するかは、予め基本設計で決めて置かねばならない。wfindの初期の実行ファイルは、両者を分離してそれぞれのデータに対応する2種の姉妹プログラムを作ることで解決した。すなわちwfind.com、wfindi.comをそれぞれ日本語MS－DOS、英語MS－DOSで利用されると想定した仕様にした。何れか一方のみを利用するユーザにとっては、このほうがかえって混乱が少ないかも知れないが、両モードを切り替えて使う場合には、2種の類似したツールを各場合に応じて区別するのは煩わしくなる。

特に支障がないならば、英語モード・日本語モードを一本化した検索ソフトにするほうが解りやすいのではないかと思われる。ただし、コードページ437や同850、etc.に基づいたDOS用の

テキストと、ウィンドウズの標準形式のANSIコードの違いは、検索者が常に意識して扱わないと、気付きにくい検索ミスを生じる重要な区別なので、両者を対等に処理するという意味での汎用化は避けるべきと判断した。従って、今回、対象データとするのはあくまで、従来のMS－DOSテキストファイル（code page 932に準拠するものを含む）ということになる(2)。

前作wfind.comやwf.exeを開発した時期に比べて、利用できるハードの性能は飛躍的に向上した。wfind.comはどちらかと言えば、機能を切り詰め不可欠のものに限定して、ファイルサイズを減少させるとともに、スピードの向上を重視した。これは8086,V30などをCPUとするパソコンで、しかもハードディスクを使用しない検索環境に土台をおいていたためであった。そのため、wfind.comはアセンブラー用のソースを作成し、またwf.exeでは画面表示の高速化を図るため、直接VRAMへの書き込みを行うルーチンを組み込んだ。アセンブラーコードは修正、追加が容易でなく、その後、指摘された一部のバグを修正した他は、ほとんど機能拡張をしなかった。また、wf.exeも機種依存の低水準処理を含むため、標準入出力を利用できない不便さが、かえって他の利点を帳消しにしてしまった。

現在、Intel 486を内蔵する機械が大半を占めるMS－DOSの世界では、ハードディスクの普及とその高速化のために、前述のようなテクニックは事実上不要化したといってよい。汎用性の大きな、まっとうなプログラムでも、実用に充分な高速性が達せられるような外的条件が生まれたのである。そこで、wfind.exe ver 2.0及びlfind.exe ver 0.8からは、既発表の自作プログラムの基本性能を受け継ぎながら、C言語を使用して新規に作成し直すことにし、コーディングがずっと楽になった。

できるだけ多くのパソコンで多くの人々が利用できるという点では、今回の検索ツールは拙稿（1994b）のWindowsソフトを凌ぐはずである。MS－DOSをOSとして利用しているがWindowsは導入していないというユーザも存在するからである。逆に、専らWindows用アプリケーションを愛用する人でも、DOSは必ず導入されているはずであるから、これらを利用することはできる。

MS－DOSがその普及度で揺るぎない地位にあることは、他のシステムでも互換性をサポートされている（される）可能性が高いということからも確認できる。最近、利用者の増加が見られているOS/2においても、そのDOSセッションで（日本語・英語のいずれでも）これらを支障なく走らせることができる。また、Macintosh上でDOSをエミュレートするSoft PCを使えば、両プログラムをまったくMS－DOS上と同様に活用できる。唯一の違いと言えば、当然ではあるが、emulationによるスピード低下であるが、実用に十分耐えるものである。

3．仕様変更と機能追加

同様な機能でありながら、いくつかの点でこれまでの拙作ソフトとは細部で異なる項目、削除

された機能、及び追加された新機能がある。これらの詳細は第５節、６節のそれぞれの検索ソフトの紹介に譲るが、両者に共通している重要な変更点を指摘しておく。

wfind.com と wf.exe が持っていた常駐ソフト Sfont への対応を廃止した。Sfont は NEC98 シリーズのパソコンで、データ上ではSTX形式という独自の区別符号方式を用いて複文字化してあるスペイン語特殊文字を、スペイン語フォントとして画面上に表示し処理するソフトであるが、今回は機種依存のこの部分を省略した。欧文特殊文字が様々な代字で表現されているか、あるいは拡張 ASCII コードが使用されていて、その字面どおりに表示されるかの、どちらかであるという択一的扱いはやめ、各欧州語を同等に扱うよう変更した。その結果、例えばSTX形式のテキストを検索した結果は、その原テキストと同じ様な表記で示されることになる。すなわち、特別な画面表示用のフォントを使用しないからといって、検索対象に適するテキストの範囲が狭まるということではない。

第２の大きな相違点はパタンマッチ関数の変更である。wfind.com / wf.exe / wwf.exe などでそれぞれのパタンマッチの条件は異なっており、詳しくは各マニュアルを参照されたいが、wfind.com を受け継いだ wfind.exe ver 2.0と比べて、最新の ver 3.01が大幅な見直しをしている点を特記しておく。

wfind.exe ver 3.01では、大・小文字の差を完全に無視する照合のみを行う。wf.exeやwwf.exe/wwfi.exe のパタンマッチでは文の冒頭のみで大小文字無視を実現していたので、文中での大文字（で始まる語、大文字のみ語）が漏れるケースが生じた。実際上、文頭でこの差を考慮しないのならば、他の部分で区分しても有意義な結果が得られないと考えられるので、この扱いを変更して、大小文字を完全に同一視することとした。さらに、前プログラムでは不十分であった、スペイン語以外の欧文アクセント文字などの特殊字の大小文字の対応をきちんと処理するよう改善した。

「語」の判定には語内文字とそうでないものの定義が不可欠である。wfind.com/wfindi.com ver.1.16 / wfindi.exe ver.2.0 と違って、新 wfind.exe 及び lfind.exe には、新たにハイフン、鈍アクセント記号及び中黒を語内文字として追加した。すなわち、－’～＾｀・の最大６種となり、これらをアルファベットの補助記号として使えば、ほぼ全てのラテン文字欧州語をカバーできると考えたためである。DOS/V の英語モード下で使用される ASCII 128以降の文字がテキストで用いられている状況でも、すべてのヨーロッパ語の字母が完備しているとは限らず、部分的に補助記号に頼る複文字化は避けられないだろう。その場合、テキスト側でどのような diacritic の割り当てを実施していようとも、１文字に６種までの下位区分が可能なので、各言語の正書法を代字化するのに十分であるはずである。

ハイフンは仏語 aprè-midi などの語を認識するために語中文字とみなさなければならないが、これをポーランド語のłをl-と表記するのに代用する手もある。また・は、カタルニア語の il・legal のように分離符号に使われる正規の句読符号だが、これをポーランド語のżを表す複文

字z•として利用することができるであろう。

ページ区切り符号の認識に関しては、従来の拙作ソフトはSTX形式のみをサポートして来た(3)。すなわち、ページ標識はそのページの第1行目の冒頭〔 〕の中に記される文字列とみなしていた。原テキストを加工して、これ以外の自己流の書式を変更することは難しくはないが、エディタ内部で数字とともに一括置換するのは面倒な面もあり、検索ソフト側でページ認識法を改善して欲しいという要望もあった(4)。そこでコマンドライン・オプション/Pを設けて検索テキストのページ標識をある程度選択できるように改めた。

複数のファイルを効率よく検索するには、既成の拙作プログラムでも、既存のファイラーのマクロ等を工夫することによって、比較的容易に実現できる。広く用いられているFD,FILMTNなどの〔Alt＋英字〕のショートカットキーに選択されたファイルを第1 argumentとしてwfind.comを起動する1行マクロを書いておけば、次々とこの検索を実行できる。また、後述のwfinds.exeを用いれば保存ファイルを指定しながら順次検索を繰り返すことも可能になる。ただ、何度もキーを押す動作は必要で、これをも省力化したい場合はやはり、ソフトに手を加えなければならず、新版wfind.exeでは当初から検索対象のマルチファイル化を採用した。

4．他ソフトとの連結

データ検索はそれだけで自己完結するのではなく、結果として得られたデータを何らかの再利用のために、修正後またはそのまま保存される必要がある。あるいは、全体を保存するか、部分セーブするか、破棄するかを決定するために、少なくとも一渡り目を通す必要がある。検索の次のステップとなるこのような作業との橋渡しがスムーズに行くかどうかは、検索ツールそのものの出来栄えと同等またはそれ以上に重要性を帯びる。

画面上から消えた出力を呼び戻すユーティリティがいくつかあり（例えばXscriptがよく知られている）、これと組み合わせればある程度の結果データを巻き戻して見ることができる。ただし、この種の画面逆スクロールのツールは限られたバッファしか持たず、数10キロ以上のデータが吐き出されると、全部を再現することが不可能になってしまう。

wfind.comの最初のバージョンでは、リダイレクトが可能であったが、検索語入力のプロンプトとコンフリクトが一部生じ、使いにくいとの声が聞かれたので、同上の派生変異プログラムとして、完全な標準入出力を基本とするフィルタ版のwfinds.exeも作成して公開した。従って、別ファイルに書き出した検索結果をエディタなどを起動してオープンすればいいのだが、このプロセスを短絡できれば能率が上がるはずである。

別のDOSプログラムとの連結は、例えば、テキストフォーマットの変換と組み合わせて使う場合に重宝するが（Cf.出口1993,pp.58－59)、最も使用頻度の高いと思われる結果データの閲覧と編集のためにその都度リダイレクトで該当ソフトと結合するのは煩雑である。もちろん、wf.exeやWindows用ソフトのwwf.exe/ wwfi.exe/wsch.exeで試みたように、これらの機能

を検索ソフト内部に実装するという解決法もある。しかし、MS－DOSにはファイル閲覧のための優れたフリーソフトもあり、またテキスト編集には自分が使い慣れたエディタを利用するのが最適と考えられるので、検索結果をスクリーンに出力するとともに、一時ファイルに落としてから、各ユーザが指定するbrowserまたはeditorを直ちに起動してみてはどうかと考えた。呼び出しプログラム名は環境変数にセットすることにし、もしセットされていなければスクリーン（標準出力）にデータを送ればよい。実際にMIELとVZ Editorを連結プログラムとしてテストを行ってみると、非常に快適に検索後処理に移行できることが判明した。ただ、エディタなどを呼び出すか否かを環境変数のみで切り替えるのはやや不便なことがわかったので、一度セットした後でも、コマンドラインオプションから無効化できるようにした。すなわち、lfind.exe/wfind.exeの両方において、/Kオプションを付ければ、現在設定中の呼び出しプログラムを起動しないようにすることができる。

lfind.exeでは、複数の単語を同時検索するのが主たる用法であるため、発見された例文のどこにどのキーワード語が存在するのか見分けやすいように、該当語を緑色表示するように改めた。出力の色付けはANSIのエスケープシークエンスを利用するので、このままリダイレクトされてfile化されてしまうと、これをエディタで加工する際に、非常に読みにくいコードが各所に混在してしまうという難点が生じた。もちろん、画面に再表示するケースを考慮すればESC sequence入りのテキストファイルがまったく不要とも言えない面もある。そこで、この点を解決するため、コマンドラインから追加のパラメータとして、出力ファイル名を指定出来るように設計した。このオプションのファイル名が、もしコマンドラインに発見されれば、色付け用のESC記号を抜き取ったネットのテキストをそこに書き出すこととする。この指定がなければ、出力は画面に向けられる。

MS－DOSはそのバージョン4.0以来、国別コード・ページ(5)を導入し、文字種とキーボードの多言語対応に乗り出し、コンピュータの国際化にやや積極的な姿勢が感じられた。新たな包括的な文字体系を作り上げるのではなく、言語別にlocalizeするという弥縫的な対策かもしれなかったが、すくなくとも、個別言語側からみれば、より正確できめ細かな表示や印刷を約束する前進には違いない。しかし、周知の通り、国内各メーカーが発売するMS－DOSはこの方向を無視して、現在なお英語のみが外国語であるという、かたくなな姿勢を崩そうとしていない。このような理由から、我が国で欧州各言語のテキストを作成するとき、それぞれの研究・教育分野でどのコード体系を採用するかについて、統一的な基準はない模様である。スペイン語関係者に関して言えば、やはり、カタカナを含んだcode page 932が多く使用されているようである。古くからのIBM系機種利用者でも英語モードのpage 437（IBM PC図形文字セット）が一般化しており、よりスペイン語に適したpage 850（多国語文字セット）はそれ程普及していないと見受けられる。

より徹底した多言語への対応を果たすためには、各国固有のcode pageで作成されたテキス

トに対する考慮がなされなければならないだろう。これらの個別対処はプログラム作成上は特に問題なく、大文字／小文字変換テーブルのみを書き換えることで、容易に実現できる。むしろ、その様な需要がどれほど存在するのか不明なため、本稿執筆の段階では、最も普及しているcode page 437をベースにした実行プログラムと、その姉妹版でcode page 850に対応するものを別個に用意するに留めた。英語版MS－DOSにおいて、一応サポートされているコードページの切り替え法が複雑で解りにくく、起動後に随時に異なるcode pageへ移動できるような仕組みになっていない点や、キーボード配列とプリンタとの連携も必要になるため、各国で予めdefault導入の初期値として設定される場合を除けば、利用しにくいのかもしれない。

なお、ギリシャ語（page 869）を除いて、個別のヨーロッパ語pageで書かれたテキストでも標準版wfind.exe/lfind.exeの検索動作は一見、正常に実行されているかのように見えるが、大小文字のマッピングが不完全なため、検索漏れが生じている可能性があるので、テキストのフォーマットを十分確認する必要がある。

II．LFIND.EXE と WFIND.EXE の概要

5．lfind.exe ver 0.95

この検索ソフトの特色、使用法はその配布用documentに詳しいので、下記にそのまま転記する：

>>

多言語対応単語リスト検索プログラム

“LFIND” ver. 0.95 (c) 1989,1994　Atumi Deguchi

英文やその他の欧文テキストで特定リストの中の単語を含む文例を同時に検索し、その出力データの一覧と編集を支援するMS－DOS用の汎用プログラムです。

◎◎特徴◎◎

1．MS－DOSの標準欧文テキスト形式（code page 437）のファイルに対し、1語以上の単語からなるリストをワンパスで高速検索し、それらを含む「文」を切り出して出力します。
2．検索結果を一覧するアプリケーション・ソフトまたは編集用に起動するエディタを指定できますので、迅速に結果処理の作業に移れます。

３．キーワードの所在をページ数、行数で報告し、各文ごとの該当事例数、テキスト全体の発見例総数も示します。該当語は緑色表示して区別されます。

４．対象テキストのページング方式に応じて、ページ区切り標識の変更を可能にするオプションを設けています。

５．行末のハイフンで分断された語形も、正しく認識します。

◎◎制限事項◎◎

１．検索対象ファイルは１バイト文字からのみなるテキストファイルで、全角日本語文字や他言語の 2 byte 文字の混在する文章は扱いません。

２．英語ウィンドウズ用標準テキストファイル（ANSI Text）は扱いません。これには拙作 wwf.exe を御利用ください。

３．連語、連句の検索はできません。この場合は拙作“wfind.com/ wfindi.com ver 1.16”を御使用下さい。

４．複数のファイルを一括して検索することはできません。この場合は拙作“wfind.exe 3.01”を御使用ください。

◎◎検索対象ファイル◎◎

べた書き文章のテキストファイルから特定の単語（関連した、あるいは、ばらばらの単語群でもよい）を含む文を見つけだすことを目的とします。市販ワープロソフトで作成された文書でもASCII テキストとして保存されたものは対象となりますが、一般に、editor 上で作成された、通常の句読符号を含んだ文章体を想定しています。

日本語 MS－DOS 上で作られた、ASCII コード127までの文字を利用する欧文（ここでは狭 ASCII と呼んでおきます）も、英語系 MS－DOS の下で使われる128－225のコードを含む欧文（拡張 ASCII として区別します）のどちらのテキストも利用できます。狭 ASCII の欧文ファイルとは、具体的な例をあげると、アクセント付きの文字や特殊な文字を何らかの補助記号で表現する、a’ a~ a` a^ a－のような複文字化した代替表記を含んでいるものを指します。補助記号は、Alphabet 文字の前後どちらにあってもよく、また２個以上重ねて使用されていても

差し支えありません。拡張ASCIIの中にこの種の複文字が混在していても構いません。

◎◎プログラムの実行◎◎

＜動作環境＞

MS－DOSが動くパソコンならどのメーカーの機種でも、どんな旧式なハードでも動作するはずです。すなわち、80x86（互換を含む）をもち、メモリーは256Kbytes程度で十分です。MS－DOS ver 3.1以上が導入されていれば、デスクトップ型・ノート型を問わず、またハードディスクのない1 Floppy Diskの機種でも問題ありません。

IBM系DOS（MS－DOS/V, PC－DOS/V, DR DOS/V, AX DOS）の日本語／英語両モードで利用できますが、config.sysにANSI.SYSを組み込んでおいてください。また、NEC,EPSON、富士通、東芝、その他の日本語MS－DOSでも使用可能です。

その他、次のシステムのMS－DOSセッションやエミュレーション下でも動作に問題が無いことを確認しました。

日本語MS－DOS Windows 3.1(NEC,EPSON,Microsoft,IBM)のDOSモード
英語MS－DOS Windows 3.1(Microsoft)のDOSモード
日本語OS/2 J2.1(IBM)の全面DOS,DOSウィンドウ
マッキントシュ SoftPC(Insignia)のDOSモード

＜起動＞

実行に必須のファイルは本体プログラム“lfind.exe”だけです。もっとも、検索の対象となる文書ファイルが予め用意されている必要があります。最も簡単な起動法は、MS－DOSのプロンプトで出ている状態で、“lfind.exe”の置かれているdirectoryで、

lfind FileName

と入力します。検索文書が“lfind.exe”と同じdirectoryにないときは、そのパス名も指定してください。

“lfind.exe”を実際に始動させる前に、環境変数EDTに検索直後に起動したいアプリケーションをセットすることができます。たとえば、出力された結果を直ちに編集したり保存するために、VZエディタを起動したいと考える場合などです。その実行プログラムがCドライブのVZというディレクトリにあると仮定すれば、

set EDT=C:¥VZ¥VZ.COM

としておきます。こうすれば、発見語がある限り自動的にこのエディタが始動して、結果データの処理へ速やかに移行することが出来ます。また、結果データ出力が１画面に収まるのは稀なため、全体を通して一読できるように何らかのFile Viewer、例えば、MIELを呼び出すのも便利な方法です。上と同様に、miel.comの存在するパスをEDT=に指定すればよいのです。

＜起動時オプション＞

プログラムを起動するときには、次のオプションを加えることができます：

lfind /F=WordList /K /P=XY FileName

オプションの意味

/F=　コマンドラインから単語のリストを入力する代わりに、単語リストを書いたファイル　WordListを読み込みます。リストのファイル形式は空白で区切られた１群の単語リストです。

/K　現在、設定されている環境変数を無視して、結果を画面に出力します。環境変数を解除するのでなく、一時的に無効にするだけです。

/P=　テキストに用いられるページ標識を定義します：＝直後の１字Xが行頭にあれば、これを新ページの開始識別子と解釈し、次の字Yを識別子の終端と解します。既定値はP/=[]ですが、まる括弧を用いるテキストであれば、/P=()を指定します。終端の記号がなく、例えば、.125のようなページ標識が使われているときは終端を改行とみなし、改行記号としてnを指定します。そこで、このような形式は/P=.nとしてください。

オプションで出力ファイル名を追加指定することも可能です；

lfind　オプション群　InFileName OutFileName

ファイル名を連記した場合は、最初のファイルが検索対象のテキストファイルとみなされ、２番目のファイルは検索結果を出力するファイル名となります。上記のようなファイル出力の場合は、カラー表示のためのESC文字は自動的に除去します。しかし、次のようにリダイレクトされたファイル出力ではESC Sequenceを維持しますので、使い分けて利用してください。

lfind　オプション群　InFileName ＞ OutFileName

＜検索語リスト入力＞

プログラムが正常に始動すると、プログラム名とバージョンが表示された後、

Enter Word List to Search ：

が出て、単語群を入力するよう促されますので、ここで検索すべき単語を半角小文字で１語以上入力してください。複数の語はスペースで区切り、最終単語の入力を終えるまでは、行末でもそのまま続けて下さい。最後に Return（Enter）キーを押し下げます。なお、何も入力せず Return を押した場合はプログラムは中断して終了します。

一度に検索することのできる語数は、各単語間の頭部に存在する共通部分の大小に左右されます。語幹が近似し語尾が異なる一連の形態のような場合は100語以上でもOKですが、関連性のない語ばかりの場合は約20～30語です。１語当たりの文字数には実質的な制限はありません。

検索語リストの入力時に語中文字として不適切な文字を用いると、プログラムは警告を発して中断しますので、起動し直してください。

＜一致条件＞

テキストの文字列と検索キーワードの比較において常に大文字・小文字の差は無視されます。本ソフトでの単語の定義は次の字種のみからなる１字以上の連続です：

1．フルファベットＡ－Ｚ，ａ－ｚ
2．code page 437 における No.128－154，No.160－165，No.225 の各文字
3．区別符号’‘~^・－

なお、区別符号のうちハイフンは行末に位置するときは語内文字としてカウントされず、照合と結果出力で常に無視されます。

＜結果表示画面＞

該当語が検出されないときは、WORD NOT FOUND!と表示します。検索語が発見されればその語を緑色で表示し、各事例毎にそのページ、頁頭からの行数を添えて、その語を含む「文」を示します。１文中に２例以上の該当例が存在するときは、その文は１度だけ出力し、文末に hit 数を書き出します。出力「文」は原文に含まれていた改行コードをすべて除去して表示します。

最後に、発見総事例数とマッチ「文」総数を知らせます。

◎◎その他の仕様◎◎

検索対象ファイルの大きさは無制限で、検索結果データを単に画面に出力する場合は、その結果データの大きさにも限度がありません。しかし、検索結果を直接ファイルに書き出すとき（2番目のファイル名を指定するか、あるいはリダイレクトするとき)、disk 上にそのファイルを作成するに十分な空き容量が必要です。

また、環境変数 EDT で連動ソフトを設定する場合は、もちろんこれらを動作させるメモリと、一時ファイル（wtemp.$$$）を作成できるスペースが必要です。

コンテキスト文の長さは最大1000文字（空白を含む）で、それ以上の場合は後部をカットします。発見位置を示す行数は概略で、その文の末尾の位置の目安となります。1文で複数の hit 事例がある場合は、最後の語の位置を示します。

文の区切りは".;:?!"の各記号を判別して行います。休止記号...は文境界標識と判断します。スペイン語などの欧文では、この処置で「文」の判別に実用上ほとんど問題が生じませんが、言語によっては、あるいは扱うテキストの文体によっては、通常の文単位に正しく区切られないケースが起こり得ます。省略語のマーカーとして使われるピリオド（例えば、Mr.Mrs.のような語）が多用される文種では、予めこれらのピリオドを他の記号に変換して迂回させる手続きが必要になるかもしれません。

◎◎プログラム概要◎◎

プログラム名：	lfind
ファイル名：	lfind.exe
Version ：	0.95
種類：	単語群一括検索プログラム
動作環境：	各社製日本語 MS－DOS 3.1以上、各社製英語系 MS－DOS, PC－DOS 4.0 以上
作者：	出　口　厚　実
ファイルサイズ：	15350 bytes
ファイル日付：	1994.08.05
開発言語：	Turbo C ver. 2.0
改版歴：	wsch10.exe 1989.5.1 NEC PC 9800 用にスペイン語 STX テキストを対象として作成
	lfind.exe v.0.8 1994.2.07　日英MS－DO 両用の汎用版に変更
	0.95　　8.05　各種コマンドオプションを追加

◎◎免責◎◎

作者はこのプログラムの仕様内容及びその信頼性を保証するものではありません。不具合いやその他お気づきの点・御要望は原作者までご連絡下されば、解決するよう努力しますが、乏しい力量のゆえにご期待にそえないこともあります。使用上の御感想などをお寄せくださった方々には改良版や関連新ソフトの情報を届けるよう努めます。

◎◎再配布◎◎

“研究教育機関に所属する個人またはグループが学術研究の目的で作成したデータ及びプログラム（コード）を広く無償で（金銭・物的等価要求を伴わずに）公開し、利用し合う”ことに賛同される方々の間では、この趣旨に従って本プログラムを自由に使用し、再配布出来ます。ただし、再配布されるときは、この document file を含めて一組でなされるものとし、そのことを原作者にご連絡下さるようお願いします。

1994.08.05　出口厚実

Nifty NBF00372

6．wfind.exe ver 3.01

この検索ソフトの特色、使用法はその配布用documentに詳しいので、下記にそのまま転記する：

>>

多言語・マルチファイル対応「文例」検索プログラム

"WFIND" ver. 3.01 (c) 1991,1994　Atusmi Deguchi

英文やその他の欧文テキストで特定の単語を含む文例を複数のテキストファイルから同時に検索し、結果データの一覧と編集を支援するMS－DOS用の汎用プログラムです。

◎◎特徴◎◎

1．MS－DOSの標準欧文テキスト形式（Code Page 437）で書かれた複数のファイルから特定種類の単語を含む「文」を切り出して出力します。

2．ファイル名に、＊記号を含むワイルド・カードが指定できますので、ファイル群をまとめた

一括検索が簡単にできます。

３．検索キーワードには＊記号によるワイルド・カードが使えます。これにより前方部分一致、後方部分一致、形態素検索も行うことが可能です。

４．検索結果を一覧するアプリケーション・ソフトまたは編集のための起動エディタを指定でき、迅速に結果処理の作業へ移れます。

５．発見語の所在をページ数、行数で示します。対象テキストのページング方式に応じて、ページ区切り標識の変更を可能にするオプションを設けています。

６．行末のハイフンで分断された語形も正しく検索します。

◎◎制限事項◎◎

１．検索対象ファイルは１バイト文字からのみなるテキストファイルで、全角日本語文字や他言語２byte文字が混在する文章は扱いません。

２．英語ウィンドウズ用標準テキストファイル（ANSIテキスト）は扱いません。これには拙作wwf.exeを御利用ください。

３．２語以上の異なる単語を同時に検索することはできません。この場合は拙作“lfind.exe”を御使用下さい。

４．出力画面でキーワードを特に色分け表示しません。このためには拙作“lfind.exe”を御利用ください。

◎◎検索対象ファイル◎◎

べた書き文章のテキストファイルから特定種類の単語を含む「文」を見つけだすことを目的とします。市販のワープロソフトで作成された文書でもASCIIテキストとして保存されたものは対象となりますが、一般に、editorで作成された、通常の句読符号を含んだ文章体テキストを想定しています。

日本語MS－DOS上で作られた、ASCIIコード127までの文字を利用する欧文（ここでは狭

ASCII と呼んでおきます）も、英語系 MS－DOS の下で使われる128－225のコードを含む欧文（拡張 ASCII として区別します）のどちらのテキストにも利用できます。狭 ASCII の欧文ファイルとは、具体的な例をあげると、アクセント付きの文字や特殊な文字を何らかの補助記号で表現する、a’ a~ a` a^ a－のような複文字化した代替表記を含むものを指します。補助記号は、Alphabet 文字の前後どちらにあってもよく、また２個以上重ねて使用されていても差し支えありません。拡張 ASCII の中にこの種の複文字が混在していても構いません。

◎◎プログラムの実行◎◎

＜動作環境＞

MS－DOS が動くパソコンならどのメーカーの機種でも、どんな旧式なハードでも動作するはずです。すなわち、80x86 の CPU をもち、256Kbyte 程度のメモリーで十分です。MS－DOS ver.3.1以上が導入されていれば、デスクトップ型・ノート型を問わず、またハードディスクのない１Floppy Disk の機種でも問題ありません。

IBM 系 DOS（MS－DOS/V, PC-DOS/V, DR DOS, AX DOS）の日本語／英語両モードで利用できます。また、NEC、EPSON、富士通、東芝、その他の日本語 MS－DOS でも使用可能です。

その他、次のシステムのMS－DOS互換ボックスやエミュレーション下でも動作に問題が無いことを確認しました。

日本語 MS－DOS Windows 3.1（NEC,EPSON,Microsoft,IBM）の DOS モード
英語 MS－DOS Windows 3.1 のDOS モード
日本語 OS/2 J2.1（IBM）の全面 DOS, DOS ウィンドウ
マッキントシュ　SoftPC（Insignia）の DOS モード

＜起動＞

実行に必須のファイルは本体プログラム“wfind.exe”だけです。もっとも、検索の対象となる文書ファイルが予め用意されていなければ役に立たないのは言うまでもありません。

最も簡単な起動法は、MS－DOS のプロンプトで出ている状態で、

wfind FileName

と入力します。ここで FileName は検索対象となる文書のファイル名ですが、wfind.exe と異なるディレクトリやドライブにあるときは、そのパス名も正しく指定してください。次のように、２つ以上のファイル名を対象にすることができます：

wfind FileName1 FileName2 Filename3 ...

また、ファイル名本体と、拡張子の表記にワイルド・カードを用いることも可能です。例えばCドライブのTEXTというdirectoryの中にある、拡張子.txtを持つファイルをすべて１度に検索したいときはC:¥text¥＊.txtと指定します。さらに、あるディレクトリ中の全ファイルを検索する場合は、単に＊.＊と書けば良いことになります。ワイルドカードを含んだファイル名をいくつか連ねることもできます。

wfind.exeを実際に起動させる前に、予め環境変数EDTに検索直後に起動したいソフト名をセットすることができます。たとえば、結果を直ちに編集したり保存するために、VZエディタを起動したいと考え、その実行プログラムがCドライブのVZというディレクトリにあると仮定すれば

set EDT=C:¥VZ¥VZ.COM

としておきます。こうすれば、発見語がある限り自動的にこのエディタが始動して、速やかに結果データの処理を行うことが出来ます。また、結果データ出力が１画面に収まるのは稀であるため、全体を通して一読できるように何らかのFile Viewer、例えば、MIELを呼び出すのも便利な方法です。上と同様に、miel.comの存在するパスをEDT=に指定すればよいのです。

<起動時オプション>

プログラムを起動するときに、次のオプションを加えることもできます：

wfind /K /P=XY FileName

オプションの意味

/K　　設定されている環境変数を無視して、結果を画面に出力します。環境変数を解除するのでなく、一時的に無効にするだけです。

/P　　テキストに用いられるページ標識を定義します：＝直後の１字Xが行頭にあれば、これを新ページの開始識別子と解釈し、次の字Yをその識別子の終端と解します。既定値は/P=[]ですが、まる括弧を用いるテキストであれば、/P=()を指定します。終端の記号がなく、例えば、.125のようなページ標識が使われているときは終端を改行とみなし、改行記号にはｎを指定します。そこで、このような形式は/P=.nとしてください。

＜検索語入力＞

プログラムが正常に始動すると、プログラム名とバージョンが表示された後、

Enter Word to Search －〉

が出て、単語を入力するよう促されますので、ここで検索すべき単語を半角小文字で１語のみ入力して、Return（Enter）キーを押してください。なお、何も入力せずReturnを押した場合はプログラムを中断して終了します。入力できる文字は語内文字（次項参照）とアステリスクに限られます。単語の頭部、尾部または両端にアステリスクを使うとその部分は０字以上の任意長の文字列と解釈されます。単語の内部にワイルドカードを用いることはできません。検索単語の長さは最大24文字までです。

＜一致条件＞

テキストの文字列と検索キーワードの比較において常に大文字・小文字の差は無視されます。本ソフトでの単語の定義は次の字種のみからなる１字以上の連続です：

１．アルファベットＡ－Ｚ，ａ－ｚ
２．code page 437 における No.128－154，No.160－165，No.225 の各文字
３．区別符号’‘~^・－

なお、区別符号のうちハイフンは行末に位置するときは語内文字としてカウントされず、照合と結果出力で常に無視されます。

キーワードに使われた文字＊は単語の左端、または右端での０個以上の任意の文字と解釈されます。例えば、検索語＊man は、テキスト側の man,woman,chairman,postman,fireman,...と一致、一方、man＊はman,manner,mania,manuscript,maneuver,...と一致します。文字列の前後に＊を配した＊man＊はman,demand,command,emancipate,...と一致します。

＜結果表示画面＞

検索語が入力されると、画面に Now Searching...と表示してテキスト検索が開始されます。検索語が発見されれば、それを含むファイル名がブラケットに囲まれて記されたのち、各事例毎にそのページ、頁頭からの行数を添えて、その語を含む「文」を提示します。１文中に２例以上の該当例が存在するときは、その文は１度だけ出力されます。従って出力文の数は、総発見事例数と同じかそれ以下になります。文例が見つからないときは、走査したファイル名のみが表示されます。出力「文」は原文に含まれていた改行コードをすべて除去します。

捜索結果を画面に出さず、直接ファイルに記録したいときは次のように出力をリダイレクトしてください：

wfind　オプション群　FileName1（FileName2...）＞　OutFileName

◎◎その他の仕様◎◎

検索対象ファイルの大きさは無制限で、検索結果データを単に画面に出力する場合は、その結果データの大きさにも限度がありません。しかし、検索結果を直接ファイルに書き出すときは、disk 上にそのファイルを作成するに十分な空き容量が必要です。

また、環境変数 EDT で連動ソフトを設定する場合は、もちろん、これらを動作させるメモリと、一時ファイル（wtemp.$$$）を作成できるスペースが必要です。

コンテキスト文の長さは最大1000文字（空白を含む）で、それ以上の場合は後部をカットします。発見位置を示す行数は概略で、その文の末尾の位置の目安となります。1文で複数の hit がある場合は、最初の語の位置を示します。

文の区切りは".;:?!"の各記号を判別して行います。休止記号....は「文」境界標識と判断します。スペイン語などの欧文では、文の判別はこの処置で実用上ほとんど問題が生じませんが、言語によっては、あるいは扱うテキストの文体によっては、通常の文単位に正しく区切られないケースが起こり得ます。省略語のマーカーとして使われるピリオド（例えば、Mr., Mrs.のような語）が多用される文種では、予めこれらのピリオドを他の記号に変換して迂回させる手続きが必要になるかもしれません。

◎◎プログラム概要◎◎

プログラム名：	wfind
ファイル名：	wfind.exe
Version ：	3.01
種類：	マルチファイル単語検索プログラム
動作環境：	各社製日本語 MS－DOS 3.1以上、各社製英語系 MS－DOS，PC－DOS 4.0以上
作者：	出　口　厚　実
ファイルサイズ：	14616 bytes
ファイル日付：	1994.08.05
開発言語：	Turbo C　ver. 2.0
改版歴：	wfind.com/wfindi.com v.1.12　1991.1.17　公開初版

wfind.exe	v.2.0	1994.2.01	日英MS－DOS両用の汎用版に変更
	3.01	8.05	各種コマンドオプションを追加

◎◎免責◎◎

作者はこのプログラムの仕様内容及びその信頼性を保証するものではありません。不具合いやその他お気づきの点・御要望は原作者までご連絡下されば、解決するよう努力しますが、乏しい力量のゆえにご期待にそえないこともあります。使用上の御感想などを報告くださった方々には改良版や関連新ソフトの情報を届けるよう努めます。

◎◎再配布◎◎

“研究教育機関に所属する個人またはグループが学術研究の目的で作成したデータ及びプログラム（コード）を広く無償で（金銭・物的等価要求を伴わずに）公開し、利用し合う”ことに賛同される方々の間では、この趣旨に従って本プログラムを自由に使用し、再配布出来ます。ただし、再配布されるときは、この document file を含めて一組でなされるものとし、そのことを原作者にご連絡下さるようお願いします。

1994.08.05　出口厚実
Nifty NBF00372

7．課題と結語

一行または前後数行のコンテキストをキーワードとともに捜し出してくれる、いわゆるKWIC形式の検索ソフトはいくつか流通している。他のテキスト処理ツールと組み合わせたパッケージの形態で配布されているものもある。実際にこれらを使用してみた経験から、足りないと思われる重要な機能をなるべく多く上記の“lfind.exe”/“wfind.exe”に盛り込んだつもりである。一方、市中のソフトで充足する機能は思い切って省いた部分が多い。今後、さらに多言語対応を実質化しなければならぬという課題が残されているが、きめ細かな option 設定を増やすには個別言語対応のプログラムに分割するほうが使い易いかもしれないという気もする。また、両プログラムに相当するソフトを他のOS上でも使えるようにするため、UNIX版、OS/2版、Macintosh版開発の必要性も痛感している。

〔注〕

1．出口（1993）では、主としてスペイン語データを対象にした、自作の汎用ファイル・コンバータを紹介した。
2．一般的な日本語 DOS/日本語 Windows の枠内で作成された欧文のデータを扱う限りではこの差を特に考えなくてもよい。なお、拙作のソフト wwf.exe/wsch.exe は日本語 Windows で動作するけれども、英語モードの Windows 用コードを生成することができるので、注意する必要がある。
3．wfindi.com/wfindi.exe ver 2.0に対しては、他の識別基準による variant を希望者に配布したことがある。
4．SED,AWK などのテキストツールを用いれば、他のページ書式を stx 形式にコンバートするのは簡単である。一例をあげれば、－PXXX（XXXは数字を示す）のページ標識を〔XXX〕に変換するためには次の SED スクリプトを書けばよい：s/__P¥([0－9]＊¥)/¥[¥1¥]/g
5．IBM DOS 5.0J/V 以降の国内売り MS－DOS で選択導入できるページは、page 850（多国語），852（スラブ語），857（トルコ語），860（ポルトガル語），861（アイスランド語），863（カナダ・フランス語），865（北欧語），869（ギリシャ語）である。

<REFERENCES>

出口厚実（1989）スペイン語テキストデータとパーソナルコンピュータの使用環境－Estudios Hispánicos 14, pp.1－13
－－　（1990a）スペイン語テキストファイルの作成とファイル型式の変換－Estudios Hispánicos 15, pp.1－15
－－　（1990b）スペイン語テキスト処理の実際：単語検索の諸問題－大阪外国語大学論集 4，pp.137－152
－－　（1991）スペイン語動詞屈折形態の同定と探索－Estudios Hispánicos 16，pp.15－28
－－　（1993）テキストデータの形式とその変換：スペイン語の場合－Estudios Hispánicos 18，pp.45－59
－－　（1994a）外国語テキスト処理に関するユーティリティの自作と活用－大阪外国語大学での情報処理・研究のあり方について,pp. 1－8
－－　（1994b）言語研究のための外国語テキストデータ検索：ウィンドウズ環境の場合－大阪外国語大学論集 11，pp.11－30

（1994. 9. 1　受理）